

International Workshop on Web Search and Data Mining (WSDM) April 29 - May 2, 2019,
Leuven, Belgium

Recommendation of collaborative filtering for a technological surveillance model using Multi-Dimension Tensor Factorization

Amelec Vilorio^{a*}, Omar Bonerge Pineda Lezama^b, Javier Reniz^c

^{a,c} Universidad de la Costa (CUC), Barranquilla 080003, Colombia

^b Universidad Tecnológica Centroamericana (UNITEC), Tegucigalpa 11101, Honduras

Abstract

Technological surveillance in research centers and universities focuses on carrying out a systematic follow-up on the development of research lines, the research leaders, the possibilities of scientific-technological collaboration, and to the knowledge of current trends from research. All these elements allow guiding the researches and supporting the scientific-technological strategy. This research proposes a model of technological surveillance supported by a recommendation system as an application that focuses on the preferences of researchers in universities and research centers. The multidimensional tensor factorization approach, based on grouping to build a recommendation system and to validate the increase in tensors, improves the accuracy of the recommendation. The experiments have been carried out in real data sets as the university and research centers. The results confirm that the dispersion issues are improved within a wider margin in both data sets. In addition, the proposed approach states that the increase in the number of dimensions shows a 7-10% improvement in accuracy and memory, which increases performance as an expert recommendation system.

© 2019 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the Conference Program Chairs.

Keywords: technological surveillance, collaborative filtering, recommendation system; academic context; research centers; multi-dimensionality; factorization tensor.

1. Introduction

Technological surveillance (TS) is a process that allows to obtain threat alerts and opportunities on the emergence of new technologies and intelligence reports with relevant information for making decisions related to organizational

* Corresponding author. Tel.: +573046238313

E-mail address: aviloria7@cuc.edu.co

technologies (Gaitán-Angulo M. et al; 2018)[1]. With the support of a collaborative platform, the TS can provide new services (Lim, H., and Kim, H. J. 2017)[2] such as the publishing of information related to the works of each researcher, the location of users and contents, the monitoring of contents, labeling of contents in a structured way, as well as obtaining consolidated information reports, which allows the detection of new projects and technologies and their evolution in a scientific community.

Despite all the advantages offered by the existence of a collaborative platform (Balasubramanian, K. et al, 2010)[3]. A for the performance of the TS process in universities and research centers, researchers do not know the people or groups of people who work on similar topics within a particular community, besides the fact that they spend a lot of time looking for pertinent information for their researches. It is particularly important to study this limitation in communities with many members (Bobadilla, J. et al, 2011)[4]. The recommendation systems allow to automatically and proactively (Adomavicius, G. and Tuzhilin, A; 2005)[5] propose opportunities for collaboration with other researchers working on similar topics, research projects, and/or similar communities.

In this research, a multidimensional generic framework is presented, based on the factorization of the tensor to address the recommendations based on the context: MD-TFCF (Multi-Dimension Tensor Factorization Collaborative Filtering). The factorization of the tensor is used since it can handle any number of contextual variables. The tensor factorization allows a flexible assimilation of contextual information by modeling the context related to the user and the products. Contextual information is considered as additional dimensions that are represented in the form of tensors. Factoring these tensors help building a unified data model that provides context-aware recommendations. So, the proposed approach allows improving the technology surveillance model for universities and research centers.

2. Technological surveillance model supported by recommendations

The term surveillance is related to the actions of observation, information gathering and its analysis, to convert scattered signals into trends and recommendations that are essential for the decision-making process (Arora, A. et al, 2016)[6]. It is necessary to obtain personalized recommendations for researchers in order to carry out surveillance. This is possible using a recommendation system or component which must be integrated into the monitoring process. Using a collaborative platform, personalized recommendations can be made to each user of the system or member of the community. The more the user interacts with the community, the better the inferences that the system makes about their tastes or preferences since the collaborative filtering is nourished by this interaction.

The systematic generation of recommendations for researchers in the academic context is a way to perform TS. Table 1 shows how each of the stages in the TS can be supported by a recommendation system, in this case, applying the Collaborative Filtering (CF) technique.

The TS model supported by recommendations based on collaborative filtering proposes the intensive use of a collaborative platform where the researcher creates his research profile and the system generates recommendations with information alerts that are possibly interesting to the user.

The Movie-Lens and Book-Crossing data sets are used, which are the two most prominent datasets accessible to the audience in the research work.

Table 1. Relation of stages in the technological surveillance with stages of the recommendations system with collaborative filtering (Harper, F. and Konstan, J; 2016)[7].

Technological surveillance	Recommendation System
Monitoring and Information Capture	Obtaining user ratings and characteristics
Information Analysis and Processing	Calculation of similarities and formation of vicinities
Intelligence	Inference of valuations and obtaining the elements to recommend
Results Dissemination	Recommendations Viewing

3. Proposed MD-TFCF Mechanism

This section presents the framework of the Multi-Dimension Tensor Factorization Collaborative Filtering (MD-TFCF) approach. The work flow of the proposed framework is shown in Fig. 1, which illustrates that the process starts from the data processing and continues to predictions according to the wishes of the users.

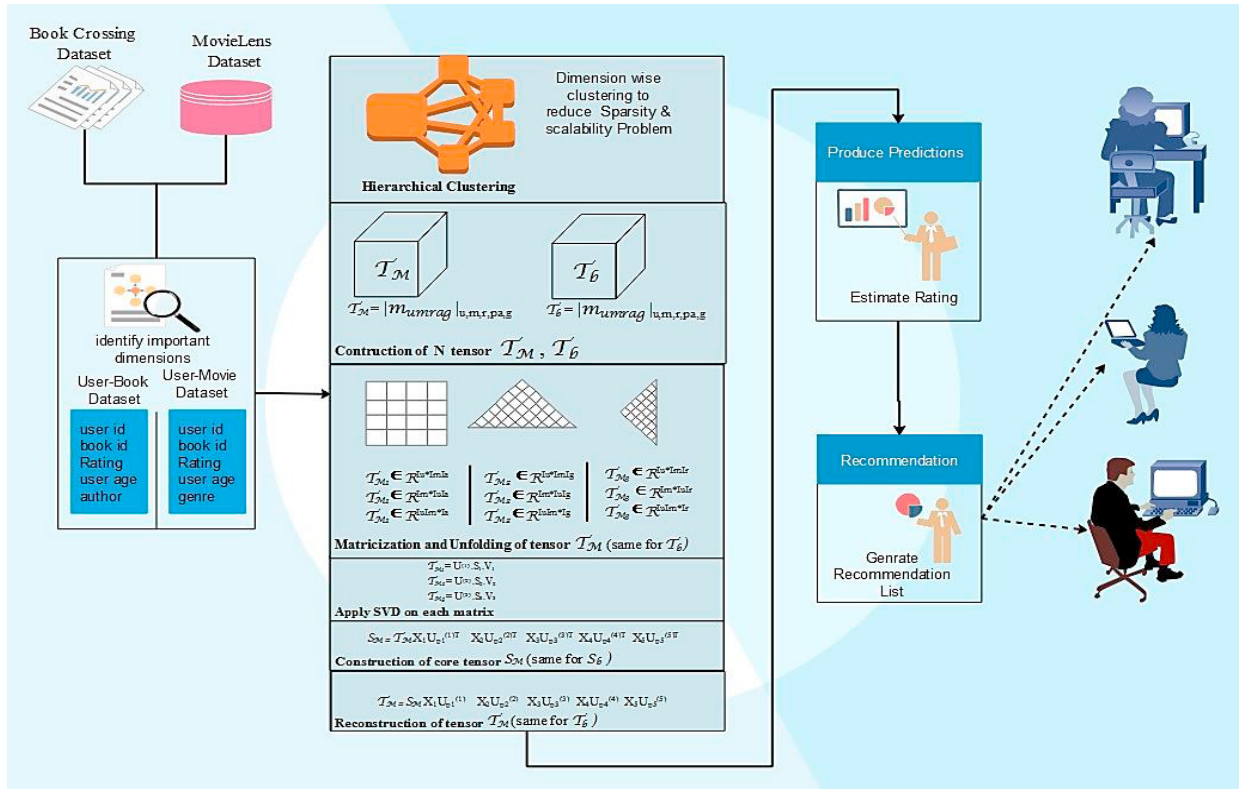


Fig. 1. Proposed Multi-Dimension Tensor Factorization Collaborative Filtering (MD-TFCF) Framework, based in (Lee, J. et al, 2016)[8]

In the consulted literature, several tensor decomposition models are available (Kolda, 2009)[9] such as PARAFAC, Tucker, Canonical, HOSVD, etc. In the proposed work, the Higher Order Singular Value Decomposition (HOSVD) Model is used to factor the tensors in matrices obtained from the user movie classification matrix. The main benefit of using HOSVD is to address the high dimensionality of the data in an effective way (Bokde, D. et al, 2015)[10], which helps to discover the relationship between users, movies/books and other contextual dimensions such as age, gender, and author.

First, the original tensor is built based on the dimensions of users, movies, ratings, age, and gender for the MovieLens dataset, and users, books, grades, age, and writer for the Book-Crossing dataset. From this point on, the tensor matrix is created conforming 28 new matrices for the Movie-Lens data set (Frolov, E. et al, 2017)[11]. More SVD is used in each newly formed matrix. Finally, the reconstruction of the central tensor and the original tensor is carried out.

The Pareto Principle which is also known as 80/20 rule is used for the verification of the predicted rating allotted through the projected MD-TFCF approach. According to the Pareto Principle, the dataset is divided and distributed evenly into training and test set in the ratio of 80% and 20% respectively. The data is evenly distributed in an 80-20

ratio so that the entire dimensions data are distributed conceptually (Braunhofer, M. and Ricci, F; 2017)[12]. The approach is experimented and assessed on cluster sets formed through hierarchical clustering approach, such as for Movie-Lens dataset. Each experiment is run 26 times (19 movie genre formed cluster + 5 user age formed cluster + 2 entire evenly distributed data) and similarly for Book-Crossing dataset. Henceforth, the prediction error is minimized using Pareto Principle as it arbitrates in evaluating the efficiency of the proposed MD-TFCF approach.

The peculiarity of a recommendation algorithm can be assessed using different forms of metrics. The suitability of the metrics used reckons on the recommendation approach, dataset, and what the recommender system will perform. Moreover, Mean Absolute Error (MAE), precision, and recall (Isinkaye, 2015)[13] are statistical measures to assess the accuracy and peculiarity of the recommendation system.

4. Experimental Results and Discussion

The TS model, supported by recommendations based on the collaborative filtering technique, was implemented in an engineering university with a flow of 5,000 visitors per day, using a collaborative platform for the management of research and the scientific work of each researcher.

In a first approach to the validation of the results, a scenario was selected within a Faculty formed by 255 researchers, the pages belonging to them in the platform (54 pages in total), and the labels the researchers use (300 labels in total), with the aim to analyze the behavior of 2 very useful metrics: the precision and the coverage of the recommended elements. In RS, precision indicates how accurate the recommendations are and measures the ability of the system to recommend only the relevant elements. On the other hand, coverage allows determining the spectrum of relevant recommendations that the system can obtain and measures the capacity of the system to recommend all relevant existing elements (Gogna, A. and Majumdar, A; 2015)[14], (Lis-Gutiérrez J. et al; 2018)[15].

Fig. 2 shows the average accuracy and average coverage of the recommendations made to the existing researchers in the scenario and their behavior according to the weight given to the labels and the elements in the calculation of the similarities among researchers. It can be seen that the accuracy and coverage of the recommendations increase as the weight of the labels increases and the weight of the elements decreases. It is pertinent to point out that the precision values do not have a noticeable increase (from 0.48 to 0.5) for the weight of the labels between 0.2 and 1.0, although these are considered significant. A difference is observed with respect to whether the labels are not used (accuracy of 0.32) and if only labels are used (precision of 0.5) (Baltrunas, L. and Ricci, F. 2014)[16].

The researchers, users of the platform (see Kamatkar S. et al 2018^a[17] and Kamatkar S. et al, 2018^b[18]), regularly receive recommendations for new pages that are added by other researchers who consider themselves similar to them. As shown by the results of the used metrics, most of the recommended pages are relevant for researchers and the vast majority of these are the ones that really interest researchers.

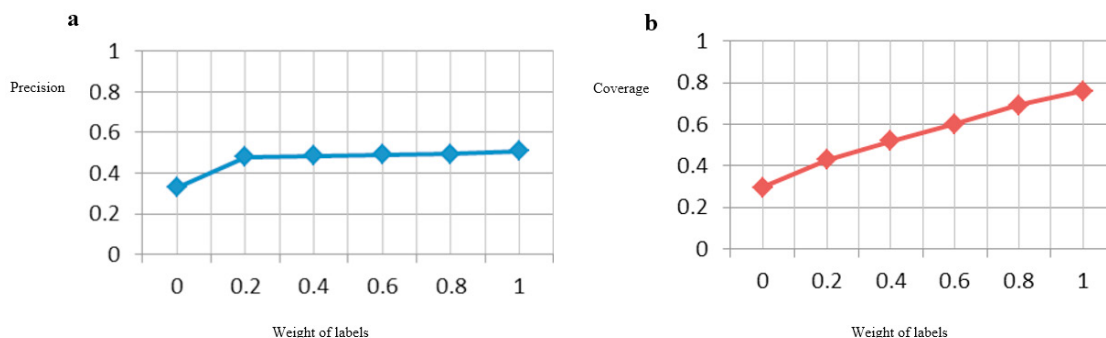


Fig. 2. Results of the accuracy and coverage of the recommendations: (a) Accuracy, (b) Coverage.

The proposed MD-TFCF approach is different from existing approaches as an integrated framework is developed in the proposed approach to unanimously represent the five dimensions. From Table 3, it can be analyzed that there are remarkable improvements in results in form of precision, recall, and mean absolute error for both Movie-Lens and Book-Crossing datasets. Table 3 infers that precision varies from 0.54 to 0.96; recall varies from 0.30 to 0.80, and mean absolute error decreases from 2.2 to 0.38 for Movie-Lens dataset, while similarly precision varies from 0.753 to 0.916, recall varies from 0.50 to 0.73, and mean absolute error decreases from 2.2 to 0.38 for Book-Crossing dataset which shows that MD-TFCF approach achieves more promising results than the traditional user-item based collaborative filtering approach. Similarly, on adding even one dimension i.e. 5-tensor approach is better than 4-tensor as accuracy in results has been improved as precision varies from 0.50 to 0.77, recall varies from 0.30 to 0.60, and mean absolute error decreases from 1.86 to 1.02. Thus, a new technique is concurrently proposed to deal with 5 dimensions and used for comparative analysis with traditional user-item based approach and with lower dimensional spaces. The proposed approach outperforms them on both real Movie-Lens and Book-Crossing datasets.

Table 2. Comparative Analysis of Higher Order Tensor with Lower Order Tensor Results

Measuring Metrics Approaches	Recall			Precision			MAE		
	Max	Min	Avg	Max	Min	Average	Max	Min	Avg
User-Item Based Neighborhood Collaborative Filtering –Movie Lens									
Normal U-I	0.5	0.1	0.23	0.5	0.05	0.38	3	1.2	2.4
Multi Dimensional Tensor Factorization Collaborative Filtering –Movie Lens									
4-order Tensor	0.6	0.30	0.50	0.77	0.50	0.65	1.86	1.02	1.46
5- order Tensor	0.80	0.30	0.65	0.96	0.54	0.81	2.2	0.38	1.06
User-Item Based Neighborhood Collaborative Filtering –Book Crossing									
Normal U-I	0.63	0.41	0.50	0.76	0.60	0.69	2.6	2	2.2
Multi Dimensional Tensor Factorization Collaborative Filtering –Book Crossing									
4-order Tensor	0.72	0.53	0.60	0.84	0.6	0.82	2.10	0.55	1.90
5-order Tensor	0.82	0.58	0.63	0.97	0.75	0.916	1.86	0.38	1.30

5. Conclusions

With the recommendation system, it is possible to increase the efficiency of the TS process since the time the researchers take to locate information relevant to them is reduced. Using the traditional search mechanisms available to the platforms, researchers need to specify the keywords or terms related to the content they wish to locate and, when the volume of information is very large (thousands of pages and users), the results of the search can contain many pages, resulting in a slow process of locating the contents of interest. With the implementation of RS, researchers do not need to use the traditional search mechanisms, since they are proactively recommended with the pages that are of interest for them.

Other techniques should be studied in more detail to obtain recommendations, such as content-based and knowledge-based techniques, as well as the different hybrid approaches that allow these techniques to be combined in order to obtain better results.

References

- [1] Gaitán-Angulo M. Amelec Viloria, Jenny-Paola Lis-Gutiérrez, Dionicio Neira, Enrique López, Ernesto Joaquín Steffens Sanabria, Claudia Patricia Fernández Castro. (2018) Influence of the Management of the Innovation in the Business Performance of the Family Business: Application to the Printing Sector in Colombia. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

- [2] Lim, H., & Kim, H. J. (2017). Item recommendation using tag emotion in social cataloging services. *Expert Systems with Applications*, 89, 179-187.
- [3] Balasubramanian, K., Kim, J., Pureskiy, A., Berry, M. W., & Park, H. (2010). A fast algorithm for nonnegative tensor factorization using block coordinate descent and an active-set-type method. *Text Mining*.
- [4] Bobadilla, J., Hernando, A., Ortega, F., & Bernal, J. (2011). A framework for collaborative filtering recommender systems. *Expert Systems with Applications*, 38(12), 14609-14623.
- [5] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6), 734-749.
- [6] Arora, A., Taneja, V., Parashar, S., & Mishra, A. (2016). Cross-domain based event recommendation using tensor factorization. *Open Computer Science*, 6(1).
- [7] Harper, F. M., & Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4), 19.
- [8] Lee, J., Lee, D., Lee, Y. C., Hwang, W. S., & Kim, S. W. (2016). Improving the accuracy of top-n recommendation using a preference model. *Information Sciences*, 348, 290-304.
- [9] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3), 455-500.
- [10] Bokde, D., Girase, S., & Mukhopadhyay, D. (2015). Matrix factorization model in collaborative filtering algorithms: A survey. *Procedia Computer Science*, 49, 136-146.
- [11] Frolov, E., & Oseledets, I. (2017). Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(3).
- [12] Braunhofer, M., & Ricci, F. (2017). Selective contextual information acquisition in travel recommender systems. *Information Technology & Tourism*, 17(1), 5-29.
- [13] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3), 261-273.
- [14] Gogna, A., & Majumdar, A. (2015). Matrix completion incorporating auxiliary information for recommender system design. *Expert Systems with Applications*, 42(14), 5789-5799.
- [15] Lis-Gutiérrez JP., Gaitán-Angulo M., Lis-Gutiérrez M., Viloria A., Cubillos J., Rodríguez-Garnica PA. (2018) Electronic and Traditional Savings Accounts in Colombia: A Spatial Agglomeration Model. In: Tan Y., Shi Y., Tang Q. (eds) *Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science*, vol 10943. Springer, Cham
- [16] Baltrunas, L., & Ricci, F. (2014). Experimental evaluation of context-dependent collaborative filtering using item splitting. *User Modeling and User-Adapted Interaction*, 24(1-2), 7-34.
- [17] Kamatkar S.J., Tayade A., Viloria A., Hernández-Chacín A. (2018)^a. Application of Classification Technique of Data Mining for Employee Management System. In: Tan Y., Shi Y., Tang Q. (eds) *Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science*, vol 10943. Springer, Cham.
- [18] Kamatkar S.J., Kamble A., Viloria A., Hernández-Fernandez L., Cali E.G. (2018)^b. Database Performance Tuning and Query Optimization. In: Tan Y., Shi Y., Tang Q. (eds) *Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science*, vol 10943. Springer, Cham